Journal of Molecular and Genetic Medicine

Liu et al., J Mol Genet Med 2018, 12:4
DOI: 10.4172/1747-0862.1000385

Mini Review Open Access

Ethnicity-Specific Reference Genome Assembly by Long-Read Sequencing

Liu Q[1], Shi L[2] and Wang K[1,3]*

[1]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[2]Guangdong-Hongkong-Macau Institute of CNS Regeneration, Ministry of Education CNS Regeneration Collaborative Joint Laboratory, Jinan University, Guangzhou, Guangdong 510632, P.R. China
[3]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

Analysis of whole genome sequencing data using a human reference genome such as GRCh38 may miss certain classes of population-specific variations. The advancement of long-read sequencing techniques has enabled efficient and effective construction of ethnically relevant reference genomes across populations. In this mini-review, we discuss recent endeavours to build ethnicity-specific reference genomes and summarize findings from these studies.

Keywords: Ethnicity-specific; Genome assembly; Precision medicine; Reference genome

Introduction

Human reference genome (Current version: GRCh38) is of fundamental importance to whole genome and exome sequencing studies, to identify single-nucleotide variants (SNVs), indels, and structural variants (SVs) that differ from the reference genome. However, GRCh38 is constructed from admixed background of contributing populations and switches from one ethnic haplotype to another at multiple places. Therefore, a universal human reference genome does not represent the ethnicity-specific haplotype diversity and may miss population-specific variations, which may be detected more efficiently using an ethnically relevant reference genome. The use of ethnicity-specific reference genome can benefit sub-population precision medicine efforts and disease variant discovery studies [1,2].

Traditionally, next-generation short-read sequencing techniques have been used to construct ethnicity-specific reference genomes. For example, one early study of ethnicity-specific genome analysis was performed on an Asian genome (YH) and an African genome [3]. Large population genome projects (such as the 1,000 Genomes Project or in Netherlands or UK) were also launched to sequence a larger number of individuals in different populations, which may facilitate the generation of ethnicity-specific reference genomes [4-7]. For example, in 2017, Maretty et al. built Denmark-specific human reference genome based on 150 individuals of 50 family trios from the Genome Denmark project [8]. Additionally, in November 2018, Sherman et al. analyzed 910 African individuals with deep sequencing data from consortium on asthma among African-ancestry populations in the Americas (CAAPA), and generated ethnicity specific contigs which cannot be aligned to the reference genome GRCh38 [9]. However, short-read based genome assembly may miss common repeats or even coding exons, and usually generated many isolated contigs that cannot be placed to chromosomes, limiting their practical use in precision medicine studies [10].

Literature Review

The rapid development of long-read sequencing techniques enables the generation of much longer raw reads and much longer assembled contigs in genome assembly studies. Thus, more and more ethnicity-specific reference genomes have become available in the past a few years. In this mini-review, we discuss recent endeavors to build ethnicity-specific reference genomes and summarize findings from these studies. For a quick reference, several existing studies on building ethnicity-specific reference genomes are summarized in Table 1.

Chaisson et al. was among the first to use SMRT (Single-Molecule Real-Time) sequencing to assemble a haploid genome using human hydatidiform mole cell line (CHM) [11]. They generated ~40X long reads data and used them to add >1.1 Mb novel sequences (i.e., sequences absent in a reference human genome) to the reference human genomes. Similarly, Huddleston, et al. sequenced CHM1 with 62X coverage and sequenced another hydatidiform mole genome (CHM13) with 66X coverage, also using SMRT long-read sequencing [12].

The NA12878 cell line is a well-studied reference cell line and is used in many genomic studies for quality assessment of sequenced reads and for benchmarking variant detection methods. In 2014, SMRT sequencing were used on NA12878 and generated 46X long reads data. *De novo* assembly on long reads generated 22,433 initial contigs with N50=906Kb (N50 is a minimum contig length where longer contigs than this value cover 50% of the assembly) [13]. After integration of genome mapping, the N50 of 202 scaffolds was improved to 31.1 Mb. Then, short-read data were combined for phasing SNVs and SVs and the generated haplotypes have > 99% consistency with trio-based results. In 2018, Oxford Nanopore long-read technique was also used to sequence NA12878 and generated ~30X long reads data [14]. *De novo* assembly of Oxford Nanopore long reads generated 2,886 contigs with N50=3 Mb, and the incorporation of 5X ultra-long reads improved N50=6.4 Mb. The detected SVs by Nanopore long reads also showed a high concordance with other previous studies using PacBio long reads and short reads. Additionally, the ultra-long reads enabled assembly and phasing of the 4 Mb major histocompatibility complex (MHC) locus in its entirety, measurement of telomere repeat length, and closure of gaps in the reference human genome assembly GRCh38 [14].

Shi et al. were the first to use long-read techniques to construct

***Corresponding author:** Dr. Kai Wang, Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia and Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, Tel: 267-425-9573; E-mail: wangk@email.chop.edu

Received December 06, 2018; **Accepted** December 27, 2018; **Published** December 31, 2018

Citation: Liu Q, Shi L, Wang K (2018) Ethnicity-Specific Reference Genome Assembly by Long-Read Sequencing. J Mol Genet Med 12: 385 doi:10.4172/1747-0862.1000385

ethnicity-specific reference genome, by assaying whole-blood sample of a diploid individual [15]. They sequenced a Chinese individual HX1 with 103X genome-wide coverage using SMRT sequencing techniques together with long-range optical mapping (a physical map by NanoChannel arrays from Bionano Genomics). Then, they constructed a Han-ethnicity genome assembly with much longer contigs than YH short-read genome assembly. The *de novo* genome assembly from long reads contain 5,843 contigs with N50 of 8.3 Mb and a total size of 2.9 Gb, and the hybrid assembly with optical mapping further improved N50 to 22.0 Mb. Detail analysis suggested that this genome assembly can fill 28.4% N-gaps in GRCh38 and contains 12.8 Mb HX1-specific novel sequences where one third were not reported in previous Asian genomes. SV analysis further suggests 9,891 deletions and 10,284 insertions in HX1, which overlap with 82.8% deletions and 66.9% insertions detected from optical mapping, and some of SVs are ethnicity-specific functional genomic elements. Long-read RNA techniques (Iso-Seq) was also used to sequence RNA samples from HX1, and the prediction from Iso-Seq data generated 58,383 high-quality consensus isoforms at 30,006 loci where 57 isoforms from 42 loci were not reported in Gencode transcript.

Later on, two parallel studies on Korean reference genome were published. In the first study, Seo et al. also took the advantages of SMRT long-read techniques together with Bionano optical mapping to sequence a Korean individual (AK1) with 101X coverage and generated Korean-specific genome assembly [16]. Their assembly resulted in 3,128 contigs with N50=17.9 Mb and 2,832 scaffolds with N50=44.8 Mb, where 8 chromosomal arms were assembled in single scaffolds [16]. Detailed analysis suggested that 1.03 Mb previous intractable sequence were added to human reference genome, and 18,210 structural variants were detected with thousands of previously un-reported breakpoints and many shared insertions in Asian population, indicating sub-population relevant patterns of reference genomes. The assembly was further phased with microfluidics-based linked reads and bacterial artificial chromosome sequencing, and generated haplotigs with N50=11.6 Mb.

Another study on Korean-specific reference genome was performed by Cho, et al. where a consensus Korean reference genome was constructed [17]. In this study, a Korean individual was sequenced using different sequencing techniques, such as short-read sequencing (~311X), SMRT long-read sequencing (~10X) and long-range optical mapping methods (240X). The assembly from short reads has 68,170 scaffolds with the length not less than 200 bp, N50=19.85 Mb and a total size of 2.92 Gb. The integration of optical mapping further improved N50 to 25.93 Mb, and then N50 was further improved to 26.08 Mb by low coverage of long reads. This genome was then integrated by common variants in 40 high-coverage short-read genomes from Korean Personal Genome Project to build a consensus Korean reference genome.

Ethnicity	Novel sequence	# of Individuals for Building Reference	Long-Read Sequencing Techniques	Reference
Utah/ Mormon	--	One	SMRT long reads, optical mapping	[13]
Utah/ Mormon	--	One	Nanopore long reads, Nanopore ultralong reads	[14]
Chinese	~12.8 Mb sequences	One	SMRT long reads, optical mapping	[15]
Korean	<1.7 Mb sequences	One	SMRT long reads, 10X Genomics linked-reads, optical mapping	[16]
Korean	8,392 SVs	One	SMRT long reads and optical mapping	[17]
Japanese	9,600 insertion sequences with ~6.2 Mb	Three	SMRT long reads	[18]
African	12,728 insertions with ~5.9 Mb	One	SMRT long reads, optical mapping	[19]
Swedish	~10 Mb sequences	Two	SMRT long reads	[20]
Puerto Rican	>20K SVs	One	SMRT long reads, Hi-C	[21]

Table 1: The summary of existing works for building ethnicity-specific reference genome.

Japanese Reference Panel was also launched to initially use short-read sequencing for a cohort of 3,552 Japanese participants for building Japanese-specific reference genome [18]. They then used SMRT sequencing techniques for 3 Japanese individuals to identify structural variants shared by Japanese. Their analysis identifies ~9,600 insertion sequences with about 6.2 Mb in size, compared with GRCh38. On 6th Jun 2017, Japanese reference genome (JRGv2) has been released after integrating *de novo* assembly from SMRT long reads. Their analysis has enabled the catalogue of Japanese-specific genetic variants.

Steinberg, et al. also used SMRT long-read technique to sequence an African individual of an Yoruban trio (NA19240) together with Illumina short reads and Bionano optical mapping techniques [19]. Their long reads data has 49X coverage, and results in a *de novo* assembly of 2.82 Gb with N50=6.1 Mb. After phasing, there are 2,230 primary contigs with N50=4.95 Mb and 9,916 haplotigs with N50=440 Kb. Bionano optical mapping data with 72X coverage can improve a scaffold N50 to 14.78 Mb for unphased contigs.

Recently, Ameur et al. sequenced two Swedish individuals using SMRT long-read techniques and optical mapping to generate Swedish specific genome assembly [20]. For one individual, 78.7X long reads data were generated, and 2.996 Gb genome with 7,166 contigs were assembled with N50=9.5 Mb. The integration with 100X Bionano optical mapping data resulted in a scaffold N50=49.8 Mb with an assembly size of 3.1 Gb. For the other individual, 77.8X long reads data were produced and the assembled genome has 7,186 contigs with a total size of 2.978 Gb and N50=8.5 Mb. Optical mapping data with 100X coverage further improved N50 to 45.4 Mb with an assembly size of 3.1 Gb. The two assemblies contain more than 10 Mb novel sequences, and 6 Mb of these novel sequences are shared with HX1. By integrating novel sequences into GRCh38, they found significant improvement of short-read alignment and variant calling: on average, removal of 10,898 false positive SNVs and additional 75,035 novel SNVs per individual.

A recent news also stated that a Puerto Rican female (HG00733, a donor to population genome projects) was sequenced with 90X using SMRT long reads, and the assembly generated 865 primary contigs with 2.89 Gb [21]. This assembly has the fewest gaps with a contig N50=27 Mb. Together with Hi-C platform, the first chromosome-scale diploid assembly (with 80% solved maternal and paternal haplotypes) of a single individual was accomplished. This assembly also contains > 20K SVs which includes Puerto Rican specific variants.

Discussion and Conclusion

In summary, the advancements of long-read sequencing techniques, including long-read sequencing from PacBio and Oxford Nanopore, linked-read sequencing from 10X Genomics and optical mapping from Bionano Genomics, have enabled efficient and effective construction of ethnicity-specific reference genome for a number of population groups. We expect that more ethnicity-specific reference genomes will be constructed and published in the future, and that these reference genomes can facilitate the analysis of next-generation sequencing data, improve accuracy of variant calling, and enable the implementation of precision medicine in diverse ethnic groups.

Acknowledgement

This study was supported by CHOP Research Institute to Kai Wang.

References

1. http://www.pacb.com/wp-content/uploads/Case-Study-Human-Improving-Precision-Medicine-Studies-in-Asia-Using-Ethnicity-Specific-Human-Reference-Genomes-and-PacBio-Long-Read-Sequencing.pdf
2. Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, et al. (2011) Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. PLoS Genet 7: e1002280.
3. Li R, Li Y, Zheng H, Luo R, Zhu H, et al. (2010) Building the sequence map of the human pan-genome. Nat Biotechnol 28: 57-63.
4. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. (2015) A global reference for human genetic variation. Nature 526: 68-74.
5. The Genome of the Netherlands Consortium (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46: 818-825.
6. Muddyman D, Smee C, Griffin H, Kaye J (2013) Implementing a successful data-management framework: the UK10K managed access model. Genome Med 5: 100.
7. Wong KHY, Levy-Sakin M, Kwok PY (2018) *De novo* human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. Nat Commun 9: 3040.
8. Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, et al. (2017) Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. Nature 548: 87-91.
9. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, et al. (2019) Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet 51: 30-35.
10. Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8: 61-65.
11. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, et al. (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517: 608-611.
12. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, et al. (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res 27: 677-685.
13. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, et al. (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 12: 780-786.
14. Jain M, Koren S, Miga KH, Quick J, Rand AC, et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36: 338-345.
15. Shi L, Guo Y, Dong C, Huddleston J, Yang H, et al. (2016) Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun 7: 12065.
16. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, et al. (2016) *De novo* assembly and phasing of a Korean human genome. Nature 538: 243-247.
17. Cho YS, Kim H, Kim HM, Jho S, Jun J, et al. (2016) An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nat Commun 7: 13637.
18. https://jrg.megabank.tohoku.ac.jp/en/
19. Steinberg KM, Graves-Lindsay T, Schneider V, Chaisson MJP, Tomlinson C, et al. (2016) High-quality assembly of an individual of Yoruban descent. bioRxiv doi: https://doi.org/10.1101/067447
20. Ameur A, Che H, Martin M, Bunikis I, Dahlberg J, et al. (2018) *De novo* assembly of two Swedish Genomes Reveals Missing Segments from the Human GRCh38 reference and improves variant calling of population-scale sequencing data. Genes 9: 486.
21. https://globenewswire.com/news-release/2018/10/08/1617872/0/en/Pacific-Biosciences-Releases-Highest-Quality-Most-Contiguous-Individual-Human-Genome-Assembly-to-Date.html